DOCUMENT RESUME

ED 226 069 TM 830 128

AUTHOR Bennett, Richard P.
TITLE Comparison of Two Analysis Techniques for the Pre-Test/Post-Test Control Group Experimental Design.
PUB DATE 27 Jan 83
NOTE 32p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Houston, TX, January 27-29, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Analysis of Covariance; Comparative Analysis; *Hypothesis Testing; *Pretests Posttests; Research Design; Sampling; Statistical Analysis; Statistical Significance
IDENTIFIERS Homogeneity of Regression; *T Test

ABSTRACT

This study examines the relative effectiveness of two means of analyzing the pre-test/post-test control group experimental design. Samples were randomly drawn from a standardized normal population and assigned to one of the four cells of the design. A set of experimental differences were induced in the post-test experimental cell. Each case was subjected to a t-test of pre-test and post-test difference scores, and an analysis of covariance (ANCOVA) using the pre-test scores as the covariate. Analysis of 200 cases in each of four sample sizes indicated that: (1) the ANCOVA is a more sensitive test having greater power to detect the induced differences but also being more susceptible to a Type I error; and (2) the findings were essentially the same when cases involving heterogeneity of regression were included and when they were excluded. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED226069

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R P. Bennett

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

COMPARISON OF TWO ANALYSIS TECHNIQUES FOR THE PRE-TEST/POST-TEST CONTROL GROUP EXPERIMENTAL DESIGN

RICHARD P. BENNETT, Ed.D

TEXAS WOMAN'S UNIVERSITY

TM 830 128

COMPARISON OF TWO ANALYSIS TECHNIQUES FOR THE PRE-TEST/POST-TEST CONTROL GROUP EXPERIMENTAL DESIGN

Summary

This study examines the relative effectiveness of two means of analyzing the pre-test/post-test control group experimental design. Samples were randomly drawn from a standardized normal population and assigned to one of the four cells of the design. A set of experimental differences were induced in the post-test experimental cell. Each case was subjected to a t-test of pre-test and post-test difference scores, and an analysis of covariance (ANCOVA) using the pre-test scores as the covariate. Analysis of 200 cases in each of four sample sizes indicated that: (1) the ANCOVA is a more sensitive test having greater power to detect the induced differences but also being more susceptible to a Type I error; and (2) the findings were essentially the same when cases involving heterogeneity of regression were included and when they were excluded.

Introduction

Huck, Cormier and Bounds (1974) state that the simplest true experimental design for behavioral research is a pre-test/post-test control group design. This design involves two groups of subjects. Both groups are tested for the dependent variable. Following application of the treatment to the experimental group, both are again tested using the same instrument. Four sets of data are produced: separate pre- and post-test scores on the experimental and control groups.

Two analysis techniques are proposed. One involves computing a difference between pre- and post-test scores for each subject. Then a t-test of independent means is applied to the sets of difference scores. As an alternative, an analysis of covariance (ANCOVA) can be employed with the pre-test data serving as the covariate (Huck, et al., 1974). The relative merit of these two techniques in terms of power and Type I error is not indicated.

The purpose of this study was to determine if there is any advantage to either of the proposed analysis techniques. Since there are differing views regarding the validity of the ANCOVA when heterogeneity of regression is present, this question was also examined.

Background

The pre-test/post-test control group design inherently lends itself to at least three analysis schemes: (1) repeated measures analysis of variance (ANOVA); (2) a simple t-test or ANOVA on the difference scores (i.e., changes between pre- and post-test scores for each subject); and (3) analysis of covariance (ANCOVA) with the pre-test score serving as the covariate. Of the three, Huck, et al. (1974, p. 247) referred students only to the latter two methods in their textbook. This decision was amplified by Huck and McLean (1975). They pointed out that the use of a repeated measures ANOVA could lead to confusion in the interpretation of results.

Huck and McLean (1975) then examined the other methods, concluding that either was acceptable with regard to interpretation. The relative power of the difference score ANOVA and the ANCOVA depended on certain conditions. Huck and McLean indicated that the two analyses would yield nearly identical results if the pre-test means for the two groups are the same. The only difference lies in the loss of a degree-of-freedom in the F-ratio denominator of the ANCOVA. The results is a slight advantage to the difference score ANOVA in terms of power. If the condition of equal pre-test means is not met, then the ANCOVA will display greater power.

The presence of equal means for the pre-test scores of the two groups is not unreasonable if these groups have been obtained by random assignment from a common subject pool. However, under these conditions, one could argue that a post-test only control group design would be sufficient and the question of preference between the difference score ANOVA and the ANCOVA would be moot. Campbell (1978) considered the analysis of only the post-test data in comparison to the other approaches and concluded that the ANCOVA was the preferred method, particularly in those instances in which control of group membership was not possible.

The use of ANCOVA in other repeated measures situations has been considered. Werts and Linn (1971) concluded that the ANCOVA was to be preferred over the repeated measures ANOVA in those situations in which the initial status of the subject was a factor affecting the changes in scores with time. The initial scores were used as the covariate. Scheifley and Schmidt (1978), in a simulation study, found that the ANCOVA had a slight advantage in error values.

In using the difference score ANOVA and the ANCOVA, certain conditions are presumed. Both techniques are based on normality of the sampling distributions and homogeneity of variance between the two groups. The ANCOVA further assumes that the regression lines for the two groups will be homogeneous and linear. Failure to meet these conditions

could distort the results of the analysis. Huck, et al. (1974, p. 66) pointed out, however, that the robustness of both techniques was such that, in the presence of heterogeneity of variance, power would not be affected if the two sample sizes are equal. This fact is supported by Scheffé (1959) who reported the results of empirical studies which demonstrated the maintenance of alpha levels.

The robustness of the ANCOVA relative to homogeneity of regression is less well established. Hamilton (1977) reported that alpha levels would be maintained in the presence of heterogeneity of regression if sample sizes are equal. Carver (1976), however, found that some variations in power might be experienced regardless of sample size, depending on the degree of heterogeneity. Hollingsworth (1980) went further, finding that heterogeneity of regression would affect the levels of alpha and power regardless of sample size and degree of heterogeneity. Consequently, it would seem appropriate to consider the degree of heterogeneity of regression when defining power levels for the ANCOVA.

Methodology

The population for this study consisted of four sets of samples randomly drawn from a standardized normal population of one million scores. The Rand Corporation (1955) tables of one million random digits was used in the selec-

tion process. Each set of samples consisted of 100 samples in one of four sample sizes (n = 10, 20, 40 or 80). Each sample was stored from lowest value to highest value.

The samples used in the study were obtained by drawing 200 sets of four numbers from 00 to 99 from the same random number table. Each group of four digits identified the samples used in a data case as described below. Sampling was without replacement to avoid data which would cause the computer to encounter singularities.

The computations were performed on a DEC-20 computer. A Fortran program was written specifically for the study. The computational format followed the steps outlined by Bruning and Kintz (1977) for an independent samples t-test with an accompanying test of homogeneity of variance and a one-way analysis of covariance (ANCOVA) with an accompanying test of homogeneity of regression. Computation of the probabilities for the t and F values used an algorithm presented by Abramowitz and Stegun (1964, p. 948). This algorithm was accurate to a level of 10^{-7}. Any probabilities reported below that value were arbitrarily treated as equal to one another when comparisons were made.

The four selected samples were assigned, in order, to be the control pre-test, control post-test, experimental pre-test and experimental post-test data sets. After considering various alternatives, pairing of the pre- and post-test data sets was accomplished by using rank-orders within the

two sets. It is understood that this method may yield higher correlations between data sets than would occur in a real setting. One of the rejected alternatives for pairing the two sets of scores was random pairing. This was rejected as being even less realistic than the chosen method. A better simulation might have been achieved by using the post-test data set, not as scores, but as randomly assigned modifiers of the pre-test scores. Each pre-test score would have been incremented or decremented by some proportion of the randomly assigned post-test "score." A fourth alternative given consideration was to use a third randomly chosen sample in some way to reorder the ranks of the post-test data set within some maximum limit for any one score. These last two alternatives had to be rejected because a limited time-frame existed for data collection.

Having established the four sets of scores and paired the pre- and post-test data sets, difference scores were computed for both control and experimental groups. The difference scores were subjected to an independent samples t-test and a test of homogeneity of variance. The original four data sets were analyzed with a one-factor ANCOVA in which the pre-test data acted as the covariate and the post-test data was the dependent variable. This was followed by a test of homogeneity of regression.

Since the original data sets were all drawn from a normal distribution, the first set of computations served to compare the two analysis techniques under the conditions of a true null hypothesis. To compare them when the null hypothesis was false, a constant score increment was added to each score in the experimental post-test data set and the tests were repeated. Six different increments were tested: .1 SE, .25 SE, .5 SE, 1.0 SE, 2.0 SE, and 3.0 SE. Consequently, for each data case, seven sets of data were generated, one for a true null hypothesis and six for a false null hypothesis with varying degrees of separation between the hypothesized and actual mean differences. The described data set was gathered at each of the four sample sizes.

All differences between the two analysis techniques were examined using the binary test (Siegel, 1954). The maximum level of significance was .05.

Results

The analysis plan consisted of examining each data case for the following. Which analysis technique performed better? Were any errors committed? Did the test meet appropriate homogeneity requirements? Under the true null hypothesis, the greater probability outcome was considered the better performance. This constituted the less significant result. Any probability below .05 was considered a Type I error. When the null hypothesis was false, lower

probability indicated greater power which constituted the better outcome. Any probabilities exceeding .05 were classified as Type II errors. Joint occurrence of probabilities below 10^{-7} were counted as equal performance. For each sample size, the 200 data cases were analyzed and summarized.

The nature of the sampling technique led one to expect the prevalence of analyses that did not meet homogeneity requirements. All comparisons of pre- and post-test scores could be expected to produce small differences between pairs. (This would be particularly true as sample size increased, thus yielding samples more representative of the normal population.) The presence of one or more unusual outlying scores could have a significant effect on the variance of a set of difference scores and, consequently, on homogeneity of variance.

Further, correlation between pre- and post-test data sets was expected to be high. Small variations in that high correlation brought about by a few extreme pairs of scores could, in turn, result in a failure of the regression lines to maintain a parallel relationship. Since it is generally accepted that failure to meet the homogeneity of variance is not a problem given equal sample sizes, failure to meet this requirement was noted but not seen as a justification to reject the data case. In contrast, the effects

of heterogeneity of regression on test validity are less well established.

Many of the data cases failed the homogeneity of regression test (Table 1). For this reason, two separate data summaries were developed. The first ignores the homogeneity requirements and summarizes all 200 data cases. The second summarizes the results of those cases in which the data displayed homogeneity of regression.

Results from all data cases

Under conditions of a true null hypothesis, the t-test produced less significant outcomes more frequently than the ANCOVA. Table 2 presents the results for those cases in which the less significant outcome did not commit a Type I error. Although significant differences between the analysis techniques are observed only for the two smaller sample sizes, the larger samples sizes indicated the same trend. If these results are augmented by those cases in which the techniques with the lesser significance also committed a Type I error (Table 3), the trend in favor of the t-test continues, with limited significance.

When attention is shifted to tests under a false null hypothesis, the greater power is displayed by the ANCOVA. (Greater power is defined as the smaller outcome probability regardless of the conclusion to accept or reject the null hypothesis.) Table 4 presents the results for those cases in which the more powerful technique did not commit

a Type II error. Only at the three smallest levels of separation between the actual and hypothesized means for the smallest sample size does the t-test yield a greater frequency of more powerful outcomes. If those cases in which the more powerful technique committed a Type II error are included (Table 5), the t-test fails to outperform the ANCOVA under any of the tested conditions. In both tables, significant differences between the two which exist are found primarily in the larger separations between means and/or larger sample sizes.

Because there were numerous Type I and Type II errors committed, their frequency is summarized. Table 6 presents the frequency of Type I errors. The occurrence of Type II errors as a function of separation level is found in Table 7. The two analysis techniques did not differ significantly on either error type regardless of the conditions of test. However, there is a consistent trend of fewer Type I errors by the t-test and fewer Type II errors from the ANCOVA.

Results with no homogeneity of regression failures

Eliminating those data cases in which the ANCOVA was deemed invalid produced a varying degree of reduction in numbers of cases analyzed dependent on sample size. The loss was consistently greater as sample size increased, ranging from 76 (38 percent) invalid cases for n = 10 to 150 (75 percent) for n = 80. Only minor changes from the earlier analysis were indicated.

Tests under the true null hypothesis again indicated that the t-test produced less significant results more often than the ANCOVA when no Type I errors occurred (Table 8). When those cases involving a Type I error were included (Table 9), this consistent trend is broken only for $n = 40$.

Tests when the null hypothesis is false produced the same pattern of results seen on the examination of all data cases. When data cases involving a Type II error are not considered (Table 10), the ANCOVA is the more powerful technique except at the three smallest separations for the smallest sample size. When the Type II error cases are included (Table 11), the trend favoring the ANCOVA continued. Only at the three smallest separations of one sample size ($n = 40$) does the t-test have a greater frequency of higher power. In none of the conditions in which the t-test showed more power were the differences between the techniques significant. Those differences which were significant all favored the ANCOVA and occurred with larger separations and/or higher sample sizes. This finding is consistent with the analysis of all data cases.

The frequency of errors of both types is indicated in Tables 12 and 13. Again, the t-test is less likely to commit a Type I error (Table 12), while the ANCOVA produced fewer Type II errors (Table 13). None of the differences between the techniques was significant.

Discussion

The results of this study must be interpreted very cautiously. As was earlier noted, the nature of the sampling procedure potentially introduced bias into the results. Further, many comparisons of the two techniques were made under a wide variety of situations. Given enough opportunities, some significant differences should occur by chance. To be more specific, comparisons of the t-test with the analysis of covariance (ANCOVA) for sensitivity were presented in Tables 2 through 5 and 8 through 11. Altogether, 112 comparisons were made. By chance, five or six of them should be significant. While there were 31 significant outcomes at p of .05 or better, cautious interpretation is still advisable. For that reason, specific differences will not be discussed. Rather, the consistent trends within the data will be examined.

The general trend of the data suggests that the ANCOVA is a more sensitive technique. In the long run, the probabilities reported by the ANCOVA were lower than those resulting from the t-test. Consequently, the ANCOVA was more likely to commit a Type I error. This is confirmed by the data of Tables 6 and 12. Although commission of Type I errors was never significantly different in number, the ANCOVA committed more errors under every condition examined except one in which it equalled the number of errors committed by the t-test.

At the same time, the ANCOVA is more likely to produce a significant result when the null hypothesis is false. This trend is supported by the data of Tables 7 and 13. With one exception, the ANCOVA committed fewer Type II errors than the t-test. Again, no one difference between the two techniques was significant.

Finding that the ANCOVA is the more sensitive test supports earlier findings. A technique in which pre-test data is used collectively to establish a baseline should be more sensitive than one in which each subject's pre-test score is used singly as that subject's baseline.

Conclusions

The results of this study support a cautious conclusion that the Analysis of Covariance (ANCOVA) is more sensitive analysis technique for the pre-test/post-test control group design. This does not necessarily make it more desirable. While showing greater power than the t-test, it also displayed a greater likelihood of committing a Type I error. The best suggestion might be to withhold any conclusion until similar data can be gathered using a more realistic method for estimating post-test scores. There may well be a performance differential between the t-test and the ANCOVA. But this difference may be small and hard to detect. Although the great majority of pre-test/post-test control group design analyses might be

unaffected by the choice of technique, finding the better technique remains a legitimate goal.

BIBLIOGRAPHY

Abramowitz, M., & Stegun, I. A., eds. Handbook of mathematical functions with formulas, graphs, and mathematical tables. Washington, D.C.: National Bureau of Standards, 1964.

Bruning, J. L., & Kintz, B. L. Computational handbook of statistics, 2nd ed. Glenview, Ill.: Scott, Foresman and Company, 1977.

Campbell, M. E. Estimation of treatment effect and measurement of change. Perceptual and Motor Skills, 1978, 46, 387-392.

Carver, D. W. The effects of nonhomogeneous regression slopes upon the power of the ANCOVA. Unpublished doctoral dissertation, University of Texas-Austin, 1976.

Hamilton, B. L. An empirical investigation of the effects of heterogeneous regression slopes in analysis of covariance. Educational and Psychological Measurement, 1977, 37, 587-601.

Hollingsworth, H. H. An analytical investigation of the effects of heterogeneous regression slopes in analysis of covariance. Educational and Psychological Measurement, 1980, 40, 611-618.

Huck, S. W., Cormier, W. H., & Bounds, W. G. Reading statistics and research. New York: Harper & Row, Publishers, 1974.

Huck, S. W., & McLean, R. A. Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. Psychological Bulletin, 1975, 82, 511-518.

Rand Corporation. A million random digits with 100,000 normal deviates. Glencoe, Ill.: The Free Press, 1955.

Scheffé, H. The analysis of variance. New York: John Wiley, 1959.

Scheifley, V. M., & Schmidt, W. H. Analysis of repeated measures data: A simulation study. Multivariate Behavioral Research, 1978, 13, 347-362.

Siegel, S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill Book Company, Inc., 1956.

Werts, C. E., & Linn, R. L. Problems with inferring treatment effects from repeated measures. Educational and Psychological Measurement, 1971, 31, 857-866.

Table 1

Frequency of failure to pass test of homogeneity

n	t-test	ANCOVA
10	55	76
20	80	93
40	103	134
80	117	150

Note: 200 cases at each sample size.

Table 2

Frequency of less significant outcome under true null hypothesis; no Type I errors present

n	t-test	ANCOVA
10	77	38**
20	49	26*
40	31	24
80	31	23

Note: 200 cases at each sample size

*$p < .01$

**$p < .001$

Table 3

Frequency of less significant outcome under true null hypothesis; including cases which yield a Type I error

n	t-test	ANCOVA
10	106	89
20	98	72*
40	81	61
30	70	44*

Note: 200 cases at each sample size

*$p < .05$

Table 4

Frequency of greater power; no Type II errors present

	Separation (standard errors)[a]					
	.1	.25	.5	1.0	2.0	3.0
n = 10						
t-test	61	57	60	53	59	54
ANCOVA	44	48	51	60	78	92
n = 20						
t-test	47	50	49	49	44	37
ANCOVA	55	51	54	56	60	51
n = 40						
t-test	34*	32*	35	35	21***	13**
ANCOVA	57	55	53	50	54	35
n = 80						
t-test	25*	30	29	22**	17	15
ANCOVA	44	47	44	45	28	24

Note: 200 cases at each sample size

[a]Separation equals actual mean minus hypothesized mean

*p<.05

**p<.01

***p<.001

Table 5

Frequency of greater power; including cases which yield a Type II error

	Separation (standard errors)[a]					
	.1	.25	.5	1.0	2.0	3.0
$\underline{n}$ = 10						
t-test	97	93	88	85**	82	64***
ANCOVA	100	104	107	127	109	113
$\underline{n}$ = 20						
t-test	74	77	79	81	68	48
ANCOVA	95	95	94	91	76	62
$\underline{n}$ = 40						
t-test	59	59	60	55	35**	22*
ANCOVA	81	79	77	74	68	42
$\underline{n}$ = 80						
t-test	43*	44*	40**	37*	25	18
ANCOVA	70	67	71	62	38	28

Note: 200 cases at each sample size

[a]Separation equals actual mean minus hypothesized mean

*$\underline{p}$<.05

**$\underline{p}$<.01

***$\underline{p}$<.001

Table 6

Frequency of Type I error under true null hypothesis

n	t-test	ANCOVA
10	86	100
20	128	131
40	144	147
80	146	155

Note: 200 cases at each sample size

Table 7

Frequency of Type II errors under varying degrees of separation

	Separation (standard errors)[a]					
	.1	.25	.5	1.0	2.0	3.0
n = 10						
t-test	105	108	97	91	60	37
ANCOVA	98	96	85	81	60	35
n = 20						
t-test	75	77	77	71	44	25
ANCOVA	67	73	71	67	41	28
n = 40						
t-test	55	58	56	50	31	20
ANCOVA	49	52	49	45	28	17
n = 80						
t-test	47	40	49	36	23	9
ANCOVA	43	35	39	32	19	8

Note: 200 cases at each sample size

[a]Separation equals actual mean minus hypothesized mean

Table 8

Frequency of less significant outcome under true null hypothesis; no Type I errors present

n	No. of Cases	t-test	ANCOVA
10	124	46	21*
20	107	28	16
40	66	8	7
80	50	11	6

*$p < .005$

Table 9

Frequency of less significant outcome under true null hypothesis; including cases which yield a Type I error

n	No. of Cases	t-test	ANCOVA
10	124	69	54
20	107	54	37
40	66	20	22
80	50	21	9*

*$p < .05$

Table 10

Frequency of greater power; no Type II errors present

	Separation (standard errors)[a]					
	.1	.25	.5	1.0	2.0	3.0
$\underline{n}$ = 10, 124 data cases						
t-test	36	36	36	31	33*	27**
ANCOVA	29	33	34	41	54	53
$\underline{n}$ = 20, 107 data cases						
t-test	23	26	25	25	25	17
ANCOVA	31	30	31	32	31	27
$\underline{n}$ = 40, 66 data cases						
t-test	13	13	12	11	5**	2*
ANCOVA	16	13	14	14	19	12
$\underline{n}$ = 80, 50 data cases						
t-test	2**	2**	3	6	7	3
ANCOVA	13	13	9	12	6	8

[a]Separation equals actual mean minus hypothesized mean

*$\underline{p}$<.05

**$\underline{p}$<.01

Table 11

Frequency of greater power; including cases which yield a Type II error

	Separation (standard errors)[a]					
	.1	.25	.5	1.0	2.0	3.0
$\underline{n}$ = 10, 124 data cases						
t-test	58	59	56	49*	50	35***
ANCOVA	65	64	65	71	68	76
$\underline{n}$ = 20, 107 data cases						
t-test	39	42	41	41	36	24
ANCOVA	54	53	53	50	42	32
$\underline{n}$ = 40, 66 data cases						
t-test	29	21	20	17	11	5*
ANCOVA	23	19	19	20	20	15
$\underline{n}$ = 80, 50 data cases						
t-test	7**	7**	7**	13	9	3
ANCOVA	23	22	21	17	11	8

[a]Separation equals actual mean minus hypothesized mean

*$\underline{p}$<.05

**$\underline{p}$<.01

***$\underline{p}$<.001

Table 12

Frequency of Type I error under true null hypothesis

n	No. of Cases	t-test	ANCOVA
10	124	58	65
20	107	65	66
40	66	49	49
80	50	36	37

Table 13

Frequency of Type II errors under varying degrees of separation

	Separation (standard errors)					
	.1	.25	.5	1.0	2.0	3.0
n = 10, 124 data cases						
t-test	64	64	60	56	36	39
ANCOVA	62	57	51	47	31	41
n = 20, 107 data cases						
t-test	42	43	43	39	26	15
ANCOVA	38	39	39	36	23	14
n = 40, 66 data cases						
t-test	15	16	15	14	8	8
ANCOVA	13	15	13	12	8	6
n = 80, 50 data cases						
t-test	17	16	18	13	6	1
ANCOVA	15	15	17	13	8	1